基于出租车司机经验的约束深度强化学习算法路径挖掘 *

黄　敏†，毛　锋，钱宇翔

(中山大学 智能工程学院 广东智能交通系统重点实验室，广州 510006)

摘　要：利用出租车司机经验，提出约束深度强化学习算法（CDRL）在线计算不同时间段内 OD 间最快路线。首先，描述了路段经验数据库（ERSD）的提取。然后，介绍了 CDRL 方法，该方法主要包括两个阶段：可选择约束路段生成和深度 Q-learning 算法，在第一阶段，生成 OD(起终点)间可选择约束路段；在第二阶段，设计深度 Q-learning 算法学习出租车司机的经验，并根据他们的出发时间计算给定 OD 间的最快路线。最后，在广州 CBD 进行了应用实验。结果表明，CDRL 方法计算在旅行时间上，优于最短路径（SR）方法，且与最快路径（FR）方法计算路径差别不大。此外，CDRL 方法在计算效率方面明显优于 FR 和 SR 方法，因此更适合 OD 间最快路径在线计算。

关键词：最快路径挖掘；路段经验数据库；经验学习；深度强化学习

中图分类号：U491　　doi: 10.19734/j.issn.1001-3695.2018.10.0810

Mining fastest route using taxi drivers' experience via constrained deep reinforcement learning

Huang Min†, Mao Feng, Qian Yuxiang

(*Guangdong Provincial Key Laboratory of Intelligent Transportation System*, *School of Intelligent System Engineering*, *Sun Yat-sen University*, *Guangzhou* 510006, *Guangdong*, *China*)

Abstract: This paper propose constrained deep reinforcement learning (CDRL) to compute the fastest route online using taxi drivers' experience in different time period. Firstly, this paper describe the extraction of experiential road segment database (ERSD) . Then CDRL method is introduced, the method is mainly comprised of two phase: bounded condition of route and deep Q-learning algorithm. In the first phase, the task is to generate alternative constrained road segments of OD pair. In the second phase, deep Q-learning algorithm is devised to learning the experience of taxi drivers, and computing the fastest route of a given OD according to their departure time. Lastly, an empirical studies is tested in CBD, Guangzhou. The results show that the routes computed by CDRL method is approximately equal to shortest route (SR) and fastest route (FR) method in travel time and route length. Furthermore, the CDRL method notably outperforms FR and SR in computing efficiency, so it is more suitable for online fastest route computation.

Key words: mining fastest route; experiential road segment database; experience learning; deep reinforcement learning

0　引言

在线搜索 OD 间最快路径已成为日常活动，并成为许多地图服务的关键功能，如谷歌和百度地图。快速行驶路径不仅节省出行者的时间、减少能源消耗，还可以缓解交通问题和保护环境，这对出行者和政府来说都很重要。良好的路径推荐系统应考虑实时交通条件和出行者驾驶行为。通常，这些信息很难提取并加入到导航系统中[1,2]。

近年来，大城市的出租车上都安装了 GPS 传感器，可以记录出租车的运动轨迹。出租车司机熟悉城市路网，他们通常根据自己的驾驶经验选择最快的路径将乘客送到目的地[3]。出租车司机选择的路径往往比地图服务软件[4]推荐的路径花费更少的旅行时间和更低的成本。 研究人员意识到，可以利用出租车司机的经验路径挖掘 OD 间最快路径，用于路径规划[5~7]。

本研究的主要目的是利用出租车司机的经验给出行者在线推荐最快的路径。即给定出行者 OD，根据他/她的出发时间推荐 OD 间最快路径。研究需要解决几个关键问题：a）在实际中，出租车司机的驾驶经验隐藏在大量的出租车 GPS 数据中[6]，应该如何从出租车历史数据中学习经验;b）路线推荐通常是实时、在线的，因此在这个系统中需要对出行者的 OD 路线计算进行快速的响应。

针对第一个问题出租车司机经验学习。常见的方法是从出租车轨迹中提取经验图，如时间依赖的地标[3, 8]、路段经验层次图[9,10]、基于网格的路径图[11]、经验路径数据库[5]以及模式感知路线图[12]。上述方法中大部分经验图的提取需要完整的 OD 路径。但由于出租车 GPS 数据的稀疏性和低采样率，很多 OD 之间不能获取足够的信息来推断给定 OD 间出租车行驶的确切路线。

对于第二个问题 OD 间路线计算，主要有两阶段路径计算法[3, 8]、约束性广度优先搜索算法[12]、最大概率积算法[13]和最短路径、最快路径算法[6]等。上述方法的计算复杂度和路网路段或者交叉口数成正相关，并且采用这些方法计算 OD 间路径需要较长的时间。

本文采用强化学习（RL）算法进行经验学习。利用出租车 GPS 数据可能不能得到 OD 间所有实际选择路径，但可以获取 OD 间所有实际选择路段。并且 GPS 数据支持每条路段的速度估计，然后可以估计路段的旅行时间，进而可以建立

收稿日期：2018-10-20；**修回日期：**2018-12-24　**基金项目：**国家自然科学基金资助项目（U1611461，11574407）；广东省科技计划项目（2016A020223006）；中央高校基本科研业务费专项资金资助项目（17lgjc42）

作者简介：黄敏（1975-），女（通信作者），广东广州人，副教授，博士，主要研究方向为交通规划与管理（huangm7@mail.sysy.edu.cn）；毛锋（1996-），男，江西赣州人，硕士研究生，主要研究方向为交通规划与管理；钱宇翔（1995-），男，广东广州人，硕士研究生，主要研究方向为交通规划与管理.

路段经验数据库(ERSD)并搜索最快路径。利用强化学习，智能体可以学习 ESRD 中隐含的出租车司机经验，然后找到最快路径。

利用强化学习的难点在于采用较少耗时的 OD 间在线最快路径计算方法。神经网络可以快速求解这类问题，将路网交叉口的状态特征输入，神经网络可以快速输出选择和交叉口相连接的各条路段的价值。

本文提出约束深度强化学习（CDRL）算法计算 OD 间最快路线。该方法主要由路径约束和深度 Q-learning 算法两个阶段组成。在第一阶段，生成 OD 间可选择约束路段。对 OD 间可选择路段进行限制隐含了出租车司机的经验，可用于智能体学习，并降低强化算法的仿真时间。在第二阶段，设计深度 Q-learning 算法学习出租车司机经验，根据他/她的出发时间在线计算给定 OD 间的最快路径。深度 Q-learning 算法包含强化学习 [14,15]和深度学习（DL）[16~18] 两个方法。利用强化学习，智能体从 GPS 数据中学习出租车司机经验，计算 OD 间最快路径，该路线可能是一条新的路径。利用深度学习，可以实时快速地计算 OD 间最快路径。

1　问题描述

本节将介绍本文中使用的一些术语，然后描述研究问题。

定义 1 路网。本文通过"节点—弧段"的方法对路网进行描述。定义有向图 $G=(E,A)$ 表示路网，其中，$E=\{e_i\}$ 为路网节点集，表示交叉口。在本研究中，用 e_S 表示起始节点，用 e_D 表示目的地节点。$A=\{a_{i,j}=\langle e_i,e_j\rangle | e_i,e_j\in N\}$ 为有向路段集。其中，$a_{i,j}$ 表示从节点 e_i 到 e_j 的有向路段。

定义 2 路段。$a_{i,j}$ 表示有向路段，用 c_{ij} 表示路段 $a_{i,j}$ 长度。定义 $\bar{v}_{ij}$ 为行驶于路段 $a_{i,j}$ 车辆的区间平均车速，$\bar{t}_{ij}$ 表示通过路段 $a_{i,j}$ 的平均行驶时间。

定义 3 路径。路径 R 由一系列连接的路段组成，i.e. $R: a_{O,1}\to a_{1,2}\to\cdots a_{n,D}$ 表示 OD 间一条路径。

定义 4 连接路段集。定义 N_i 为节点 e_i 的下一节点集，$L_i=\{a_{i,j}|e_j\in N_i\}$ 表示和节点 e_i 相连接的路段集。

定义 5 转向规则（TRI）。定义 $T=\{t_j=\langle a_{i,j},e_j,a_{j,k}\rangle | e_i\in E\}$ 表示交叉口转向规则：$a_{i,j}$ 表示当前所在路段，v_j 表示当前所在交叉口，$a_{j,k}$ 表示下一路段。

定义 6 路段经验数据库（ERSD）。ERSD 记录每天不同时间段内路段速度、旅行时间等信息，信息是从出租车 GPS 数据中提取的,后面将详细介绍。

研究问题定义：给定出行者 OD 及出发时间，利用从出租车 GPS 数据中提取的经验路段数据库(ESRD)和交叉口转向规则(TRI)，在线计算 OD 间最快路径。

2　路段经验数据库提取

本章将介绍经验路段数据库（ERSD）的提取，然后描述路段平均车速估计，以及旅行时间估计。

2.1　路段经验数据库

良好的路径推荐系统应考虑实时交通条件和出行者驾驶行为。路段行驶时间变化性至少体现在两方面：

a）时变性。路段上的交通流量随时间变化，进而影响路段行驶时间。例如，道路可能在高峰时段变得拥挤，在非高峰时段通畅行驶。

b）空间变化性。不同的道路具有不同的时变交通模式。例如，一些道路即使在高峰时段也通畅行驶。但是有些道路的高峰时段可能会持续一整天[3]。交通模式随时间变化可能导致出租车司机作出不同的路线选择，因此本文构建基于时间段的经验路段数据库。根据广州道路交通的特性[9]，将每天分为早高峰时段(7:00~9:00)、晚高峰时段(7:00~8:00)和其他非高峰时段。

本文每条 ERSD 数据将记录路段在早高峰时段、晚高峰时段、其他非高峰时段的路段平均车速，以及路段在在早高峰时段、晚高峰时段、其他非高峰时段的旅行时间。

2.2　路段平均速度估计

将出租车的 GPS 位置信息匹配到路网地图上，GPS 点在路段的分布情况可分为以下两类，类型 1:同一辆车在某段路上留下两个或两个以上的 GPS 点；类型 2:同一辆车在某段路上只留下一个 GPS 点。

类型 1 的 GPS 点分布如图 1 所示。某车辆在该路段留下了两个以上的 GPS 点，本文计算首定位点和末定位点到下游交叉口的位置 s_i^1, s_i^2，以及时间计算两点的距离和两点的时间差 $t_i^2-t_i^1$，从而得到这辆车的平均速度。求和某时段经过该路段的所有车辆的距离和时间差，可得到该路段 $a_{i,j}$ 的路段速度 $\bar{v}_1^{ij}$。

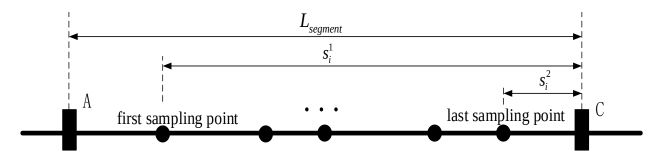


图 1　类型 1：同一辆车在某段路上留下两个或两个以上 GPS 点

Fig. 1　Type1: taxi leaves two or more than two GPS points on a certain road segment

类型 2 的 GPS 点分布如图 2 所示。对于车辆只在目标路段 $a_{i,j}$ 上留下一个 GPS 点，本文通过把路段分成若干个△L，将位于△L 范围内的所有 GPS 点的点速度的平均值作为△L 范围内的平均速度 $v_{\Delta L}^{i}$，得到若干个区间速度，最后将求出所有区间速度的平均值，将其作为路段的速度 $\bar{v}_2^{ij}$。

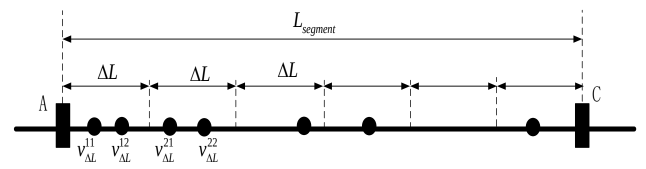


图 2　类型 2：同一辆车在某段路上只留下一个 GPS 点

Fig. 2　Type2: taxi leaves only one GPS point on a certain road segment

假设出租车 GPS 数据中类型 1 GPS 点的比例为 ω_1，则类型 2 GPS 点比例为 $1-\omega_1$。路段 $a_{i,j}$ 的路段平均速度可由式(1)计算：

$$\bar{v}^{ij}=\omega_1\bar{v}_1^{ij}+(1-\omega_1)\bar{v}_2^{ij} \tag{1}$$

2.3　路段旅行时间估计

根据定义 2 及路段平均速度 $\bar{v}_{ij}$，路段 $a_{i,j}$ 旅行时间可由式(2)计算：

$$\bar{t}_{ij}=\frac{c_{ij}}{\bar{v}_{ij}} \tag{2}$$

3　路径学习和计算

本章提出用于在线计算 OD 间最快路径的约束深度强化学习（CDRL）方法。首先介绍了 OD 间可选择约束路段生成，然后描述用于学习 OD 间最快路径的强化学习算法，最后介绍了用于 OD 间最快路径学习和在线计算的 CDRL 方法。

3.1　约束路段生成

如研究[10]所述，出租车司机经验隐藏在 GPS 数据中，这种经验是从成千上万次出行中积累起来的，隐含着他们对道路网络和真实交通状况的熟悉程度。出租车司机通常根据驾驶经验选择最快的路径，尽量减少行驶路径上发生交通拥堵情况，将乘客送达目的地。由于出租车 GPS 数据的稀疏性和低采样率，很多 OD 之间不能获取足够的信息来推断出租车行驶的确切路线。但出租车 GPS 数据足够大，可以获知 OD 间所有行驶路段的交通数据，出租车司机的经验同样隐藏在 OD 间行驶的路段中。因此，可以通过学习 OD 间行驶路段来学习出租车司机经验。

对于最优路径选择问题，需要生成 OD 间路径选择集。Ramming [19]和 Frejinger 等人[20]指出，通常很难生成 OD 间包含所有实际选择路径的路径选择集。在实际中，出行者往往只选择 OD 间若干条路径行驶，即 OD 间可选择路径存在一定约束。为了避免生成的路径选择集遗漏重要的路径，并且生成路径集满足约束，本文采用数据挖掘方法。对于每个 OD 对，选择足够长的采样时间段 T，提取该时间段 GPS 数据中出租车所有行驶路径，利用行驶路径可获得 OD 对间可选行驶路段集 $s_{O,D}$，然后通过搜索由可选行驶路段集 $s_{O,D}$ 组成的路网来获得 OD 间所有可能选择路径。

对于每个 OD 对，选择足够长的采样时间段 T。如果出行者在 T 时间段内，所有实际选择的路径中总共包含 n 条路段，就可以认为这 n 条路段组成的路段集合，可作为该 OD 对可选行驶路段集 $s_{O,D}$ 。

由于出租车 GPS 数据足够大，获得 OD 间所有实际行驶路段的数据是容易的。该方法的优点是可以生成 OD 间所有实际选择行驶路段，获得的选择路径集不会遗漏实际中重要的路径。OD 间可选行驶路段集隐含了出租车司机的经验，可用于智能体学习，降低强化学习算法的仿真时间。同样，该方法存在一个缺点，需要一个很长的采样时间段 T 来获取一个稳定的 OD 间所有实际选择路段集 $s_{O,D}$ 。

3.2　强化学习

本文中出行者根据导航和驾驶经验，选择从一个交叉口行驶到另一个交叉口，并从环境中获得收益，目标是选择一条 OD 间由可选择路段连接组成的效益最大路径。该过程类似于 MDP 过程，可用强化学习算法解决。强化学习算法包含几个重要部分：状态空间 S 、动作空间 A 、奖励函数 r 、折减系数 γ 。

出行者的状态 $s \in S$ 表示出行者在路网中所在交叉口 $e_i \in E$ 。在交叉口 e_i 的动作集 $A(e_i)$ 表示和交叉口 e_i 相连接的路段 $a_{i,j}$ ， $a_{i,j} \in L_i$ 。奖励函数 $r(e_i,a_{i,i+1})$ 表示出行者在交叉口 e_i 选择路段 $a_{i,i+1}$ 的收益。本文研究目标是计算 OD 间最快路径，因而可用路段 $a_{i,i+1}$ 旅行时间的负值 $-\bar{t}_{ij}$ 表示动作奖励，如式(3)所示。

$$r(e_i,a_{ii+1})=-\bar{t}_{ii+1} \tag{3}$$

折减系数 $\gamma \in (0,1)$ 表示当前动作选择对未来的影响程度，γ 越接近 1，表明当前状态选择动作对未来影响程度越大。

该问题的解决方法是寻找一个策略 π ，策略表示状态到动作的一个映射。在本研究，策略 π 表示出行者在交叉口选择的行驶路段，表示为 $\pi(e_i)=a_{i,i+1}$ 。出行者执行策略 π 和环境交互得到由状态、动作、奖励组成的回合（episode），表示为 $h_{i:K}=e_i,a_{i,i+1},r(e_i,a_{i,i+1}),e_{i+1},a_{i+1,i+2},r(e_{i+1},a_{i+1,i+2})\cdots,e_K$ ， e_K 表示出行者最终所在交叉口。定义 $G(h_{i:K})$ 表示 episode 累积折减收益，表示为 $G(h_{i:K})=\sum_{k=i}^{K-1}\gamma^{k-i}r(e_k,a_{k,k+1})$ 。

出行者需要找到策略 $\pi(e_i)$,使得 episode 累积折减收益 $G(h_{i:K})$ 最大。假设采用贪婪策略，总是选择使 $G(h_{i:K})$ 最大的动作。特别地，定义 $Q(e_i,a_{i,i+1})$（Q 值）为出行者在交叉口 e_i 选择路段 $a_{i,i+1}$ 最大累积折减收益，$Q(e_i,a_{i,i+1})=r(e_i,a_{i,i+1})+\gamma \max G(h_{i+1:K})$，则

$$\pi(e_i)=\arg\max_{a_{i,i+1}\in A(e_i)} Q(e_i,a_{i,i+1}) \tag{4}$$

若 $Q(e_i,a_{i,i+1})$ 值已知，就可以通过式（4）求解最优策略。本研究中 $Q(e_i,a_{i,i+1})$ 值未知，但式(4)满足 Bellman 方程性质，因而可以通过从后往前迭代估计 $Q(e_i,a_{i,i+1})$ 值，如式(5)所示。

$$Q(e_i,a_{i,i+1})=r(e_i,a_{i,i+1})+\gamma \max_{a_{i+1,i+2}\in A(e_{i+1})} Q(e_{i+1},a_{i+1,i+2}) \tag{5}$$

强化学习算法需要不断迭代更新 $Q(e_i,a_{i,i+1})$ 值，在最开始学习中 $Q(e_i,a_{i,i+1})$ 估计值与实际值会相差很大，但随着每次更新迭代，估计值变得越来越准确。

3.3　约束深度强化学习

传统的强化学习求解，Q 值估计通常使用一个 Q 值表或一个函数近似器实现 [21,22]。然而，如果最优路径规划问题的状态空间很大，存储这么一个表将很消耗时间和内存，而函数近似方法不能解决实时最优路径规划问题。

本研究采用深度 Q-learning 算法来估计 Q 值。在深度 Q-learning 算法中，使用深度神经网络作为状态映射到 Q 值的函数近似器。使用神经网络作为近似器，可以解决拥有更大的、连续的状态空间的最优路径规划问题[23]。本研究使用的神经网络中，将出行者起始地所在交叉口的状态特征作为输入，输出起始地到目的地的旅行时间。

本文将出行者所在交叉口 e_i 的状态特征表示为 $s(e_i)=[x_i,y_i,x_D,y_D]$， x_i,y_i 分别表示交叉口 e_i 的经度、纬度，x_D,y_D 表示目的地交叉口 e_i 的经度、纬度。交叉口 e_i 到目的地的旅行时间用 $Q(s(e_i))$ 表示，将出行者所在交叉口 e_i 的状态特征 $s(e_i)$ 输入本文设计的神经网络，就能得到交叉口 e_i 到目的地的旅行时间 $Q(s(e_i))$ 。

在深度 Q-learning 算法中，出行者执行策略 π 和环境交互得到由状态、动作、奖励组成的 episode，$h_{i:K}=e_i,a_{i,i+1},r(e_i,a_{i,i+1}),e_{i+1},a_{i+1,i+2},r(e_{i+1},a_{i+1,i+2})\cdots,e_K$，若出行者最终所在交叉口为目的地所在交叉口，即 $e_K=e_D$，则称该 episode 为 success episode。

深度 Q-learning 算法中，当智能体完成一次 success episode，Q 值发生更新，将 episode 中出行者在交叉口 e_i 的每次选择记录表示为 $\langle s(e_i),a_{i,i+1},r(e_i,a_{i,i+1}),s(e_{i+1})\rangle$， 存储于集合 episode memory E 中。当智能体每次完成 success episode，计算 success episode 中每个交叉口 e_i 到目的地的累积折减收益（旅行时间负值）$G(h_{i:K})$。定义 $N=\{\langle s(e_i),q(e_i)\rangle \mid e_i\in E, q(e_i)=\min G(h_{i:K})\}$ 表示 node memory N，该集合中二元组 $\langle s(e_i),q(e_i)\rangle$ 存储交叉口 e_i 的状态特征及交叉口到目的地的最短旅行时间。

本研究采用的深度 Q-learning 算法，神经网络的训练可以通过最小化交叉口 e_i 到目的地的最短旅行时间 $q(e_i)$ 和交叉口 e_i 到目的地的旅行时间估计值 $Q(s(e_i))$ 误差平方和，即

$$L(\theta)=\sum_{\langle s(e_i),q(e_i)\rangle\in N}\left(q(e_i)-Q(s(e_i),\theta)\right)^2 \tag{6}$$

当出行者完成一定 success episode 得到策略后，将面临 explore-exploit 困境:即选择当前最优策略，或者继续探索寻找可能的更优策略。本研究采用 ε 贪婪策略，以 ε 概率选择当前最佳策略，1-ε 概率随机选择策略。

基于深度 Q-learning 算法，结合 OD 间可选约束路段集，

本研究提出了 CDRL 算法，算法具体步骤如下表伪代码描述。

Algorithm1 CDRL 算法

输入：路网 $G=(E,A)$；OD 间可选约束路段集 $s_{O,D}$；交叉口转向规则 TRI。

1 初始化 node memory D

2 初始化动作价值函数 Q 及神经网络权重系数 θ

for episode = 1, M do

初始化 episode

for step = 1, K do

3 在交叉口 e_i，满足交叉口转向规则 TRI 时，采用 ε 贪婪策略选择和交叉口相连的路段 $a_{i,i+1}\in L_i$

4 将选择记录 $\langle e_i,a_{i,i+1},r(e_i,a_{i,i+1})\rangle$ 加入 episode 及记录 $\langle s(e_i),a_{i,i+1},r(e_i,a_{i,i+1}),s(e_{i+1})\rangle$ 存储于 episode memory E

if $e_{i+1}==e_D$ then

break

end for

5 计算 success episode 中每个交叉口 e_i 到目的地的累积折减收益 $G(h_{i:K})$，并更新 node memory N

6 使用梯度下降更新 θ，以最小化 $(q(e_i)-Q(s(e_i),\theta))^2$

end for

输出：策略 $a_{i,i+1}=\pi(e_i)$。

4　实例应用

在本章中，为了评估 CDRL 方法的性能，算法应用于广州市出租车司机 OD 路径选择实例研究中，并将结果与基于 Dijkstra 算法的最快路径（FR）和最短路径（SR）方法计算的结果进行比较。

4.1　实验数据及预处理

本文选择广州市天河区 CBD 作为实例研究区域，天河区 CBD 道路网络有 345 个路段和 202 个交叉口。本文使用的出租车 GPS 数据是广州市 1800 多辆出租车从 2015 年 6 月 1 日至 2015 年 6 月 21 日，超过 4.74 亿条出租车 GPS 记录，并基于此数据库提取了路段经验数据库（ERSD）。

图 3 所示是广州天河区 CBD 早高峰路段平均速度估计结果。路段上数字表示的是该路段早高峰路段平均速度估计值。从图 3 可以看出，广州 CBD 在早高峰时段出租车行驶速度缓慢，因为早高峰时段大量城市居民要前往 CBD 上班。CBD 区域内横向、纵向的几条主干道行驶速度明显高于区域内的支路，因为广州 CBD 内大量公司位于支路旁边，出租车在主干路上能正常行驶，乘客一般在支路上下车。

图 4 所示是广州天河区 CBD 早高峰路段旅行时间估计结果。路段上数字表示的是该路段早高峰路段旅行时间估计值。

选取 2015 年 6 月 1 日至 6 月 14 日获取的路段经验数据库作为模型的训练集，选取 2015 年 6 月 15 日至 21 日获得的路段经验数据库作为模型的验证集。

4.2　CDRL 算法训练

基于从出租车 GPS 数据中提取的路段经验数据库，采用 CDRL 方法学习出租车司机经验。训练数据集是 2015 年 6 月 1 日至 2015 年 6 月 14 日提取的路段经验数据库。

在 CDRL 模型中，输入广州天河区 CBD 路网 $G=(E,A)$，CBD 区域内若干个 OD 对的经纬度及可选约束路段集 $s_{O,D}$，以及交叉口转向规则 TRI，模型的训练通过最小化 OD 间最短旅行时间和实际旅行时间估计值的误差平方和。

图 5 显示了在早高峰时段、晚高峰时段、其他非高峰时段 CDRL 方法的神经网络损失函数收敛曲线。可以看到，CDRL 方法在早高峰时段迭代到约 2 700 次收敛，在晚高峰时段迭代约 1 800 次收敛，在其他非高峰时段迭代约 2 300 次时收敛。另外，三个时间段，在训练阶段初期，神经网络损失函数收敛曲线振荡明显，因为强化学习算法在初期学习阶段对 OD 间最短旅行时间估计误差很大，随着循环次数增多，最短时间估计值越来越准确。

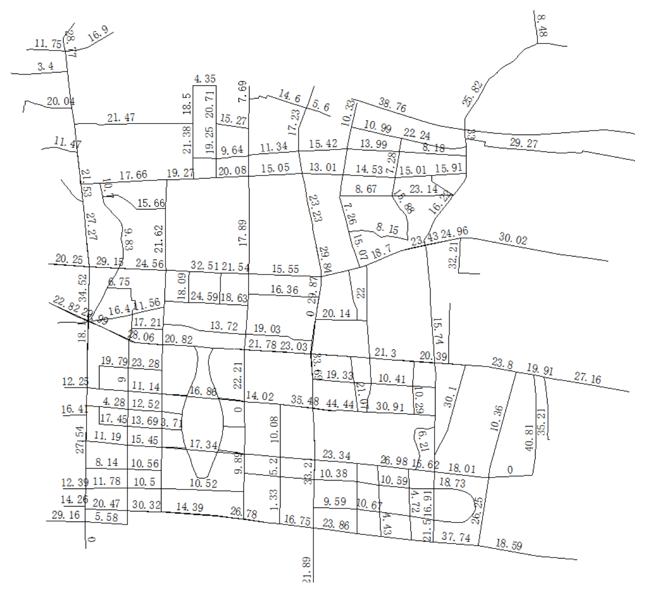


图 3　CBD 早高峰路段平均速度估计值

Fig. 3　Speed estimation of road segment in morning peak hours in CBD

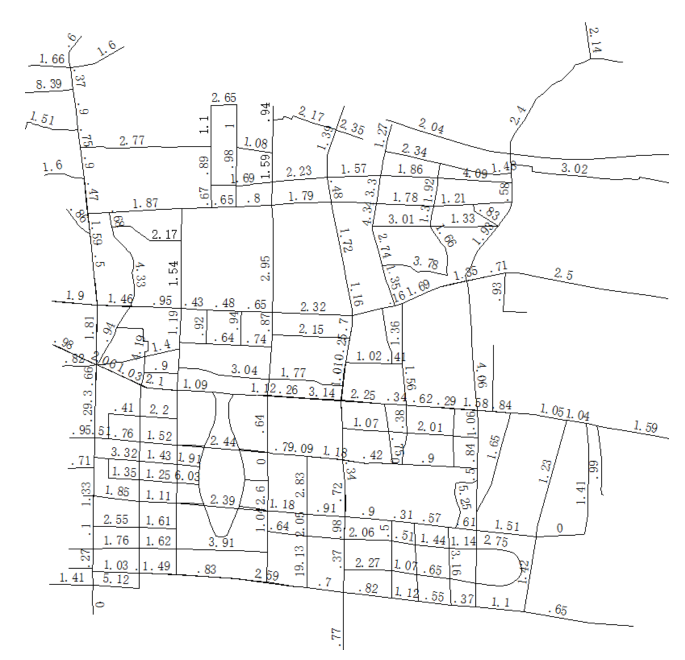


图 4　CBD 早高峰路段旅行时间估计值

Fig. 4　Travel time estimation of road segment in morning peak hours in CBD

神经网络损失函数反映旅行时间预测的误差。CDRL 方法将出行者所在交叉口和目的地的经纬度输入神经网络，输出交叉口到目的地的旅行时间，若路网所有路段的交通状况相同，则神经网络损失函数的值可以接近零。图 5 中其他非高峰时段预测误差最小，因为该时段 CBD 路段交通流密度接近；早高峰预测误差最大，因为早高峰 CBD 一部分路段可能拥堵，而有些路段却通畅，路段流量差别大；晚高峰时段预测误差居中，因为晚高峰 CBD 大部分路段都很拥堵。

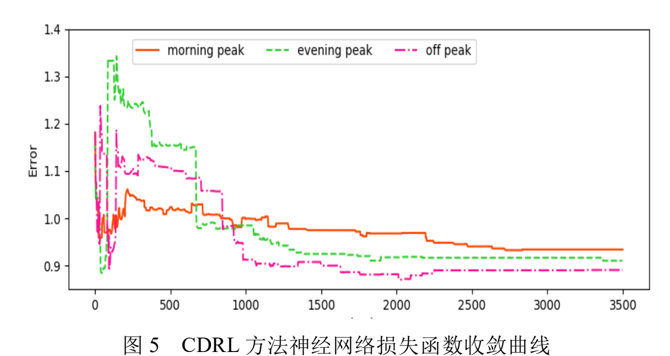


图 5 CDRL 方法神经网络损失函数收敛曲线

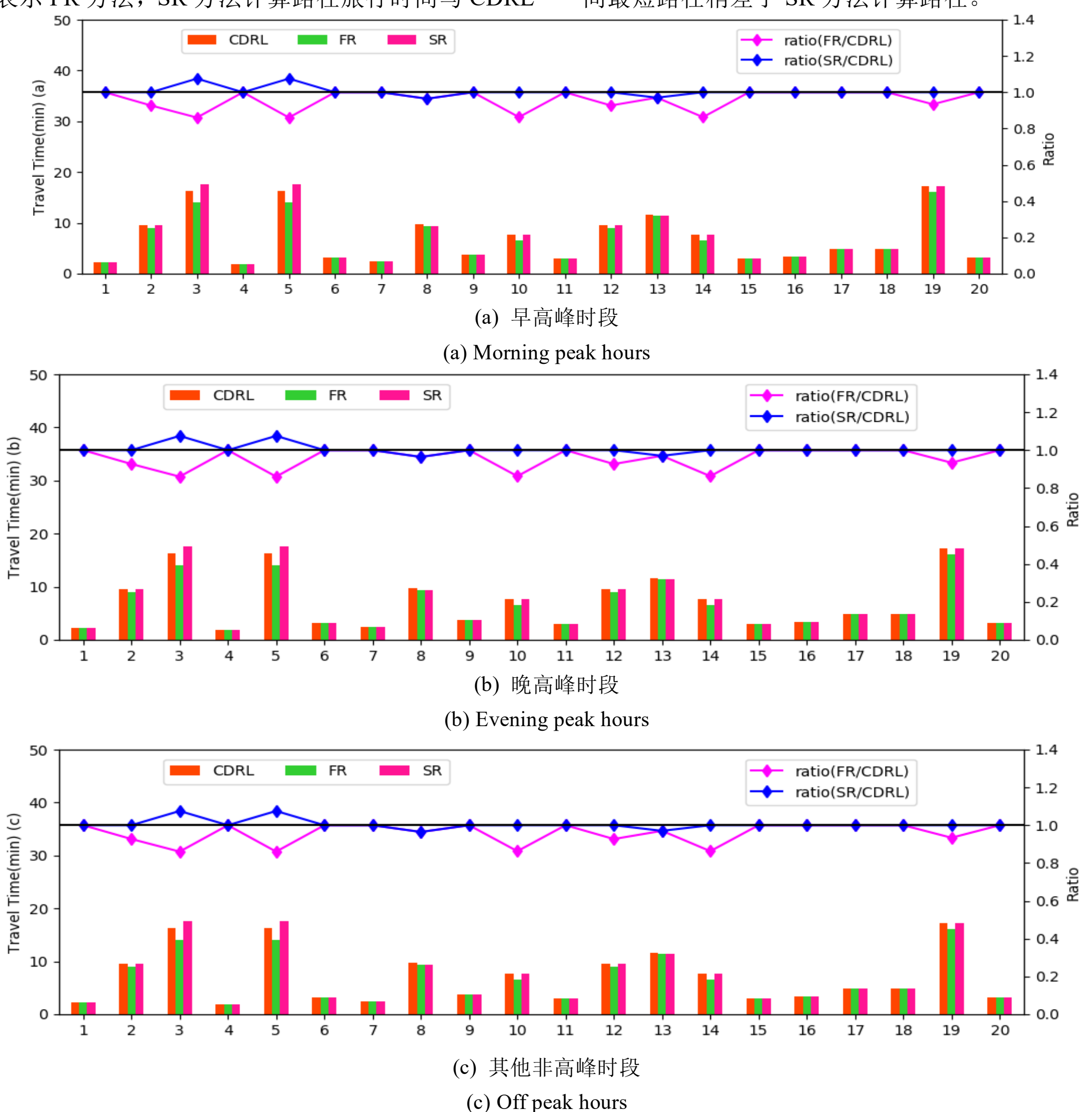
Fig. 5 Convergence curve of loss function by CDRL method

4.3 旅行时间对比

本文随机选择广州 CBD 区域内 20 个 OD 对进行实验，验证数据集是 2015 年 6 月 15 日至 2015 年 6 月 21 日提取的路段经验数据库，分别应用 CDRL 方法、FR 方法和 SR 方法。FR 方法中，使用路段旅行时间作为路段价值，然后采用 Dijkstra 算法计算 OD 间旅行时间最短路径。SR 算法中，使用路段长度作为路段价值，然后采用 Dijkstra 算法计算 OD 间长度最短路径。

图 6 显示了采用三种方法计算 OD 间路径的旅行时间对比。图中条形块的高度表示计算路径的旅行时间，图右侧纵坐标刻度表示 FR 方法，SR 方法计算路径旅行时间与 CDRL 方法计算路径旅行时间的比值。从图 6 可以看出，采用 FR 方法计算的 OD 间路径，在早高峰时间段，大部分路径的旅行时间小于或等于 CDRL 方法的结果，但差距不大，FR / CDRL 比值在 0.8~1.0 间。因为 FR 和 CDRL 方法都倾向于选择旅行时间最短的路线，但由于 CDRL 训练回合数不够，训练网络未至最优，晚高峰时段和非高峰时段实验结果和早高峰时段类似。采用 SR 方法和 CDRL 方法计算的路径，早高峰时段，SR / CDRL 的比值在 0.95~1.1 间，基本接近于 1。但 CDRL 计算的约 80%路径旅行时间比 SR 方法短。因为 CDRL 算法，通过学习出租车司机经验，会选择当前策略下旅行时间最短路径，所以所选路径旅行时间优于 SR 方法所选路径。

图 7 显示在划分的三个时段内采用三种方法计算 OD 间路径的路径总旅行时间对比。可以看出，在三个时间段内，FR 方法计算的路径总旅行时间最短，因为 FR 算法选择 OD 间旅行时间最短路径。其次则为 CDRL 方法计算路径，在晚高峰时段和非高峰时段，CDRL 方法计算路径的总旅行时间都小于 SR 方法计算路径的总旅行时间。SR 方法计算路径的总旅行时间最长，只在早高峰时段计算路径的总旅行时间小于 CDRL 方法，因为早高峰时段 CBD 区域道路流量差别大，CDRL 方法计算路径误差相对较大，因而当前策略下旅行时间最短路径稍差于 SR 方法计算路径。



(a) 早高峰时段

(a) Morning peak hours

(b) 晚高峰时段

(b) Evening peak hours

(c) 其他非高峰时段

(c) Off peak hours

图 6 CDRL, FR and SR 方法计算路径旅行时间对比

Fig. 6 Route travel time comparison for CDRL, FR and SR

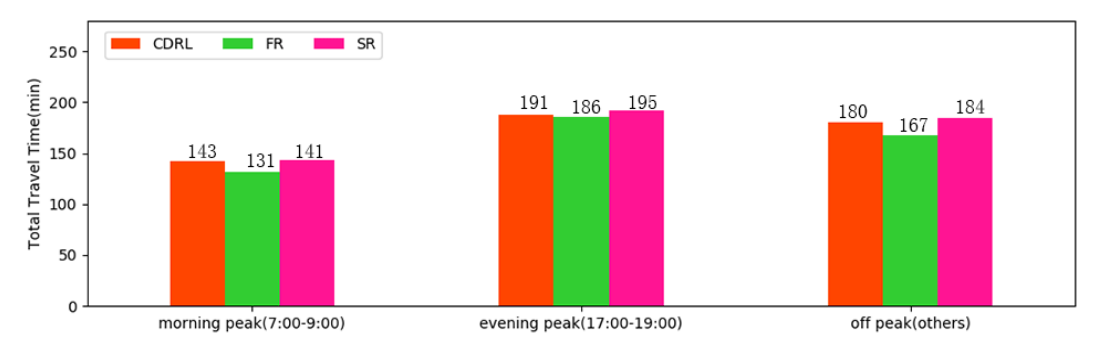


图 7　CDRL, FR and SR 方法计算路径的总旅行时间对比

Fig. 7　Total travel time comparison for CDRL, FR and SR

4.4　计算效率对比

本文采用 CDRL、FR 和 SR 三种方法计算 OD 间路径，实验仿真是在 CPU: AMD FX(tm)-4130 Quad-Core Processor、8 GB 内存、3.8 GHz 主频的计算机上实现，程序采用 Python 编程语言实现。选取 2015 年 6 月 15 日至 2015 年 6 月 21 日提取的路段经验数据库作为验证数据集，使用 4.3 节中的 20 个 OD 对进行实验，本文使用算法运行时间来评价三种方法计算效率。

CDRL 方法中，输入 20 个 OD 对的起终节点所在交叉口的经纬度，路网及路段经验数据库，然后使用训练好的模型计算 20 个 OD 对间最快路径，记录算法运行时间。FR 和 SR 方法中，将 20 个 OD 对及路网输入模型，然后计算各个 OD 间的最快路径，并记录算法运行时间。

表 1 表示在划分的 dg 个时段内，采用三种方法计算 OD 间路径的计算时间差异。可以看出，在三个时段，采用 SR 方法计算路径的计算时间是相同的，因为它们没有考虑实际交通信息。三个时间段的 SR 的计算时间都是相同的，因为它们没有考虑实际的交通情况。CDRL 方法，FR 方法在高峰时段的总计算时间高于非高峰时段，因为在高峰时段，算法需要搜索更大的路网节点空间以获得 OD 间路线。且在早高峰时段和晚高峰时段的总计算时间是差别不大。

从表 1 可得到，CDRL 方法的总计算时间最短，因为训练好的用于计算旅行时间神经网络，可以快速地计算各个交叉口到目的地的旅行时。其次为 SR 方法计算路径所需总时间，FR 的总计算时间最长。 此外，CDRL 方法的计算时间远远小于 SR 和 FR 方法。 因此，它更适合在线 OD 间最快路径计算。

表 1　CDRL, FR and SR 方法总计算时间对比

Table 1　Total calculate time for CDRL、FR and SR

	早高峰时段	晚高峰时段	其他非高峰时段
CDRL	1.981s	1.972s	1.717s
FR	6.304s	6.243s	6.075s
SR	4.999s	4.999s	4.999s

5　结束语

随着人工智能的发展，强化学习在最优路径规划方面的应用引起学者关注。与传统的路径规划方法不同，在 RL 算法中，智能体在交叉口，通过选择路段与环境进行交互，得到来自环境对选择路段效益的价值反馈，并根据环境给出的奖励调整其动作，从而选择 OD 间最优的路径。

本文提出通过学习出租车司机经验来在线计算 OD 间最快路径的 CDRL 方法。 实证研究表明，该方法计算的路径在旅行时间方面优于 SR 方法，与 FR 方法差别不大。此外，CDRL 方法在计算效率方面明显优于 FR 和 SR 方法，因此更适合在线计算 OD 间最快路径。本文认为，这种经验学习，深度强化学习方法与并行智能交通系统[24,25]的结合具有巨大潜力，可能改变下一代 ITS 发展历程。

本文提出的的方法也存在一些缺陷和不足。该方法的神经网络训练时是回合制(episode)更新，效率低，后期将设计更优的神经网络结构用于算法训练。

参考文献:

[1] Yang Lin, Kwan M P, Pan Xiaofang, *et al.* Scalable space-time trajectory cube for path-finding: a study using big taxi trajectory data [J]. Transportation Research Part B Methodological, 2017, 101: 1-27.

[2] Tang Luliang, Chang Xiaomeng. The knowledge modeling and route planning based on Taxi'experience [J]. Acta Geodaetica Et Cartographica Sinica, 2010, 39 (4): 404-409.

[3] Yuan Jing, Zheng Yu, Zhang Chengyang, *et al.* T-drive: driving directions based on taxi trajectories [C]// Proc of Sigspatial International Conference on Advances in Geographic Information Systems. [S.l.]:ACM Press, 2010: 99-108.

[4] Wong M O C K. Modelling uncertainty in traffic and transportation systems [J]. Transportmetrica, 2009, 1 (1): 1-3.

[5] He Zhaocheng, Chen Kaiyin, Chen Xinyu. A collaborative method for route discovery using Taxi drivers'experience and Preferences [J]. IEEE Trans on Intelligent Transportation Systems, 2018, 19(8): 2505-2514.

[6] Zhang Jindong, Meng Weibin, Liu Qiangqiang, *et al.* Efficient vehicles path planning algorithm based on taxi GPS big data [J]. Optik-International Journal for Light and Electron Optics, 2016, 127 (5): 2579-2585.

[7] Zheng Jiangchuan, Ni L M. Modeling heterogeneous routing decisions in trajectories for driving experience learning [C]// Proc of ACM International Joint Conference on Pervasive and Ubiquitous Computing. [S.l.]:ACM Press, 2014: 951-961.

[8] Yuan Jing, Zheng Yu, Xie Xing, *et al.* T-drive: enhancing driving directions with Taxi drivers [J]. IEEE Trans on Knowledge & Data Engineering, 2012, 25 (1): 220-232.

[9] Hu Jihua, Huang Ze, Deng Jun. A hierarchical path planning method using the experience of Taxi drivers [J]. Procedia-Social and Behavioral Sciences, 2013, 96: 1898-1909.

[10] Li Qingquan, Zeng Zhe, Zhang Tong, *et al.* Path-finding through flexible hierarchical road networks: an experiential approach using taxi trajectory data [J]. International Journal of Applied Earth Observation & Geoinformation, 2011, 13 (1): 110-119.

[11] Wei Lingyin, Zheng Yu, Peng W C. Constructing popular routes from uncertain trajectories [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]:ACM

Press, 2012: 195-203.
[12] Wei Lingyin, Chang K P, Peng W C. Discovering pattern-aware routes from trajectories [J]. Distributed & Parallel Databases, 2015, 33 (2): 1-26.
[13] Chen Zaiben, Shen Hengtao, Zhou Xiaofang. Discovering popular routes from trajectories [C]/ /Proc of IEEE International Conference on Data Engineering. [S.l.]:IEEE Computer Society, 2011: 900-911.
[14] Sutton R, Barto A. Reinforcement learning: an introduction (2nd Edition) [M]. [S.l.]:MIT Press, 2017.
[15] Gu Shixiang, Holly E, Lillicrap T, *et al.* Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates [C]// Proc of IEEE International Conference on Robotics & Automation. 2017.
[16] Chen Yuanfang, Shu Lei, Wang Lei. Poster abstract: traffic flow prediction with big data: a deep learning based time series model [C]// Proc of Computer Communications Workshops. 2017.
[17] Lv Yisheng, Duan Yanjie, Kang Wenen, *et al.* Traffic flow prediction with big data: a deep learning approach [J]. IEEE Trans on Intelligent Transportation Systems, 2015, 16 (2): 865-873.
[18] Polson N G, Sokolov V O. Deep learning for short-term traffic flow prediction [J]. Transportation Research Part C Emerging Technologies, 2017, 79: 1-17.
[19] Ramming M S. Network knowledge and route choice[D]. Cambridge:Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 2002.
[20] Frejinger E, Bierlaire M. Capturing correlation with subnetworks in route choice models[J]. Transportation Research Part B, 2007, 41: 363-378.
[21] El-Tantawy S, Abdulhai B, Abdelgawad H. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto[J]. IEEE Trans on Intelligent Transportation Systems, 2013, 14 (3): 1140-1150.
[22] Ozan C, Baskan O, Haldenbilen S, *et al*. A modified reinforcement learning algorithm for solving coordinated signalized networks[J]. Transportation Research, Part C: Emerging Technologies, 2015, 54: 40-55.
[23] Mnih V, Kavukcuoglu K, Silver D, *et al.* Human-level control through deep reinforcement learning[J]. Nature, 2015: 518 (7540): 529-533.
[24] Li Li, Ding Wen. Parallel systems for traffic control: a rethinking[J].. IEEE Trans on Intelligent Transportation Systems, 2015, 17 (4): 1179-1182.
[25] Chen Cheng, Wang Feiyue. A self-organizing neuro-fuzzy network based on first order effect sensitivity analysis[J]. Neurocomputing, 2013, 118: 21-32.